ISSN: 2349-6002

Real-Time AI-Based Multi-Channel Spam Detection System Using Transformer Models

Abhinav Pandey[1], Shraddha Bisht[2], Srashti Pal[3], Deepak Kumar Pathak[4], Siddharth Pandey[5]
[1,2,3,4,5]*Department of Computer Applications, Invertis University, Bareilly, India*
doi.org/10.64643/IJIRTV12I11-207249-459

***Abstract*—Unsolicited and malicious digital messages have become a serious concern for communication security, particularly because spam, phishing, and fraud attempts now appear across SMS, email, social media, and instant messaging platforms. This paper presents a real-time AI-based spam detection framework using transformer models for contextual message classification. The study compares traditional machine learning classifiers with a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model using the SMS Spam Collection dataset. Classical models, including Naïve Bayes, Logistic Regression, Random Forest, and Support Vector Machine, produced strong accuracy values between 97% and 98%. However, the fine-tuned BERT model achieved better overall performance, with 99% accuracy, 0.97 spam recall, 0.98 spam F1-score, two false positives, and five false negatives. The confusion matrix further indicates that BERT is more effective in identifying short and context-dependent spam messages because it captures semantic meaning rather than relying only on surface-level word patterns. The findings suggest that transformer-based architectures are suitable for real-time spam filtering, especially in modern multi-channel environments where attackers use personalized and AI-assisted phishing content.**

***Index Terms*—BERT, Spam Detection, Multi-Channel Security, Transformer Models, F1-Score, Phishing.**

I. INTRODUCTION

India has experienced a sharp rise in cybercrime and digital fraud in recent years. Official records show that nearly 1,402,809 cybercrime cases were reported in 2021, reflecting the increasing pressure of phishing, spam messages, and online fraud on individuals, organizations, and financial systems [1], [2]. Phishing remained one of the common forms of cyber-enabled fraud during this period, as attackers frequently used unsolicited messages to mislead users and obtain sensitive information. By 2023, the threat environment had become more severe, with social engineering, identity misuse, and payment-related fraud becoming more visible in the Indian digital ecosystem [3]. Some states also reported strong regional growth in cybercrime. For example, Gujarat recorded a 41% increase in reported cybercrime cases in 2023 compared with 2022, showing that spam-driven fraud and identity-based attacks were not limited to national-level statistics but were also visible at the state level [4].

The scale of the problem continued to expand by 2025. The Indian Computer Emergency Response Team reported handling more than 2.94 million cyber incidents across the country, including phishing, malware activity, and organized spam campaigns [5]. Financial losses were also reported to exceed ₹19,812 crore, with phishing-led UPI scams, social engineering attacks, and AI-assisted frauds contributing substantially to the total loss [6]. Maharashtra, Karnataka, Uttar Pradesh, and Telangana were among the major affected states, indicating that the risk is concentrated in digitally active regions as well as in areas where mobile-based financial transactions are widely used.

The rapid growth of spam and phishing in India is closely related to structural and technological changes. More than 850 million active internet users have widened the digital attack surface, while the use of smartphones and UPI-based payment platforms has created new opportunities for cybercriminals. Automated tools, botnets, and bulk messaging systems have made large-scale spam campaigns easier to execute. At the same time, generative artificial intelligence has enabled attackers to prepare messages that appear grammatically correct, personalized, and context-aware. The fragmented nature of digital communication, covering SMS, email, social media,

and messaging platforms, further increases the number of entry points available to attackers. These developments show the limitation of conventional spam filtering methods and justify the need for a real-time AI-driven multi-channel spam detection system that can understand context and support timely threat mitigation.

II. LITERATURE REVIEW

Spam detection has gradually evolved from simple rule-based systems to advanced machine learning and transformer-based approaches. In the present digital environment, users communicate through email, SMS, social media platforms, and instant messaging applications. With the expansion of these channels, spam messages, phishing links, and fraudulent content have also increased. A recent review of AI-based phishing and spam detection methods examined more than 130 studies published between 2020 and early 2025, focusing mainly on machine learning and deep learning techniques [7]. Similarly, Goswami et al. [8] discussed the use of artificial intelligence for identifying online spam, showing that AI-based spam filtering has become an active research area in both academic and applied cybersecurity domains.

Early spam detection systems mainly depended on keyword matching, blacklist databases, and manually prepared heuristic rules. These methods were computationally simple and useful for detecting repeated spam patterns, but they were not flexible enough to handle modified spellings, changed sentence structures, or newly generated phishing content. As spam campaigns became more automated and adaptive, rule-based systems became less suitable for large-scale deployment [9]. Later, supervised machine learning techniques such as Naïve Bayes, Support Vector Machine, Random Forest, and Logistic Regression were used for spam classification [10]–[12]. These models generally used feature extraction methods such as Bag-of-Words, TF-IDF, and n-grams. Studies during 2015–2020 reported accuracy values between 85% and 95% on standard email datasets, indicating that traditional machine learning could perform reasonably well under controlled conditions [12].

Despite their usefulness, traditional machine learning models face limitations in modern spam detection. Most of these methods depend on surface-level word frequency and statistical patterns rather than deeper sentence meaning. For example, the sentences “Your bank account statement is available” and “Your bank account is locked. Click here to verify” may contain similar terms such as bank and account, but the second sentence carries urgency and possible manipulation. A purely statistical model may not properly understand this difference. Naïve Bayes also assumes conditional independence among words [10], which is not realistic for natural language because meaning depends on relationships between words. Similarly, models such as SVM do not directly consider word order or syntactic structure [11]. These limitations reduce their ability to detect subtle phishing intent, tone manipulation, and adversarial rephrasing, especially in AI-generated spam messages.

Deep learning models were later introduced to address some of these contextual limitations. Long Short-Term Memory networks improved sequential understanding, while Convolutional Neural Networks supported automatic feature extraction from text. These approaches generally performed better than traditional classifiers, but they still faced challenges related to long-range dependencies, computational cost, and real-time scalability. Transformer architectures offered a major improvement in natural language processing because of their self-attention mechanism, which allows the model to examine relationships among all words in a sentence simultaneously [21]. BERT further strengthened this approach by learning bidirectional context, meaning that it reads text from both left-to-right and right-to-left directions [20]. This allows the model to capture phishing intent, urgency, and semantic cues more effectively than earlier models.

Transformer models are especially useful in present spam detection because modern phishing messages are often created with advanced language models and appear fluent, personalized, and contextually rich [26]. Instead of depending only on keyword frequency, BERT-based models can distinguish between legitimate alerts and malicious messages by examining tone, intent, urgency, and word relationships. Recent studies from 2021 to 2025 have generally reported transformer-based models with accuracy above 97%, higher F1-scores, lower false positive rates, and better generalization across domains. However, a research gap still remains. Many existing studies focus only on single-channel datasets

and do not fully address real-time deployment, multi-channel integration, or continuous adaptation against AI-generated phishing. Therefore, a real-time transformer-based spam detection framework is needed to improve contextual intelligence, scalability, and robustness against modern spam campaigns.

III. METHODOLOGY

This study follows an experimental and model-based design to develop a real-time spam detection framework for multiple communication channels, including SMS, email, and social media. The proposed approach treats spam detection as a binary text classification problem in which each message is classified as spam or ham. The framework uses both traditional machine learning models and a transformer-based BERT model so that the performance of statistical feature-based methods can be compared with contextual deep learning.

The dataset used for the broader framework is designed to represent multi-channel communication. It includes publicly available email spam data, SMS spam data, and social media comment spam data. The Enron Email Dataset is a widely used public corpus containing nearly 500,000 internal emails exchanged by senior executives of Enron Corporation during the 2001 federal investigation [13]. It provides sender, receiver, subject, date, message body, and folder-level organization, making it useful for natural language processing, fraud analysis, and email communication research. The SMS component is based on the Kaggle SMS Spam Collection dataset, which contains 5,574 English text messages labelled as spam or ham [14]. After cleaning, nearly 5,169 usable records are commonly available for spam detection experiments. Social media comment datasets from platforms such as YouTube and Twitter/X are also useful for spam detection because they contain short, informal, and noisy user-generated text [15]. These datasets may include fields such as comment text, tweet ID, hashtags, retweets, timestamps, and video IDs, and they are commonly used in NLP, sentiment analysis, and user behaviour studies.

Before model training, raw text is preprocessed to remove noise and improve consistency. Non-informative content such as HTML tags, URLs, special symbols, and irrelevant punctuation is removed. Text is converted to lowercase to maintain uniformity, and tokenization is applied to split messages into meaningful word units or n-grams using tools such as NLTK [16]. Stop words are removed when they do not contribute to classification, and stemming or lemmatization is applied to reduce words to their root forms [17], [18]. Since spam messages are often short, informal, and noisy, abbreviations and slang terms may be expanded where required. Emojis and emoticons are not always removed because they can carry sentiment or intent; instead, they may be converted into textual meaning. Numbers are either removed when irrelevant or replaced with a number token when they are important in financial spam or phishing content.

Real-time and multi-channel deployment requires preprocessing that is accurate but not overly slow. Lightweight Python libraries such as NLTK, spaCy, scikit-learn, and regular expressions are used for cleaning and text handling. The preprocessing strategy also changes according to the channel. SMS messages are usually short and compact, email and social media posts may be longer and noisier, and image-based spam may require optical character recognition before text classification. For contextual understanding, transformer models such as BERT are used because they can capture semantic meaning in multilingual, informal, and complex spam messages.

Feature extraction converts cleaned text into numerical representations suitable for classification. Traditional models use Term Frequency–Inverse Document Frequency, which assigns higher weight to terms that are frequent in a particular message but relatively rare in the overall corpus [19]. This helps identify spam-related terms such as “lottery,” “claim,” “urgent,” or “offer,” while reducing the influence of common filler words. However, TF-IDF vectors are sparse and do not capture word context. Therefore, this study also uses transformer embeddings generated through BERT. In this approach, each message is tokenized and passed through the transformer encoder, and the contextual representation of the [CLS] token is used as a fixed-length feature vector for classification [20]. Unlike TF-IDF, BERT embeddings are dense and context-aware, allowing the model to detect hidden spam patterns in short and noisy messages.

For model building, baseline classifiers such as Naïve Bayes, Logistic Regression, Support Vector Machine, and Random Forest are trained using TF-IDF features

[10]–[12], [22]. These models provide a reference point for evaluating the added value of transformer-based learning. The BERT model is then fine-tuned for binary classification by passing contextual embeddings through a classification head. The architecture includes tokenized input, transformer encoder layers, a dense hidden layer with ReLU activation, and a sigmoid output layer for spam probability estimation. This combined experimental design helps examine both statistical word patterns and deeper contextual meaning in spam detection.

For general training, the dataset may be divided into 70% training, 15% validation, and 15% testing subsets. The model is trained using labelled data, Binary Cross-Entropy loss, and the Adam optimizer [23], [24]. Hyperparameters such as learning rate, batch size, and number of epochs are tuned using validation data to reduce overfitting. Since spam datasets are often imbalanced, class weighting, oversampling of the minority class, or undersampling of the majority class may be used to improve spam recall [25]. For transformer-based models, the input sequence length is standardized, for example to 128 tokens, and attention masks are used to handle variable-length messages efficiently.

Model performance is evaluated using accuracy, precision, recall, F1-score, and confusion matrix. Accuracy measures overall correctness, while precision indicates how many messages predicted as spam are actually spam. Recall is particularly important because it measures how many actual spam messages are correctly detected. F1-score provides a balanced view of precision and recall, and the confusion matrix shows true positives, true negatives, false positives, and false negatives. In spam detection, both precision and recall are important because false positives may block legitimate messages, while false negatives allow harmful spam to reach users [25].

IV. EXPERIMENTAL SETUP

The experimental evaluation compares traditional machine learning classifiers with a transformer-based BERT model for SMS spam detection. The experiment uses the publicly available SMS Spam Collection dataset, which contains 5,574 English SMS messages labelled as spam or ham [14]. Spam messages are encoded as 1, while legitimate ham messages are encoded as 0. The dataset is divided into 80% training data and 20% validation data. The validation set contains 1,115 messages and is used to assess model performance on unseen samples.

Before training, text normalization and tokenization are performed. For traditional machine learning models, TF-IDF vectorization is applied with 5,000 maximum features and an n-gram range of (1,2) [19]. This allows the models to use both individual words and short word combinations. For the transformer-based model, the BERT tokenizer is used with padding and truncation up to a maximum sequence length of 128 tokens [20]. The baseline classifiers include Multinomial Naïve Bayes, Logistic Regression, Linear Support Vector Machine, and Random Forest [9], [11], [12]. All baseline models are trained using TF-IDF features.

The transformer model uses the bert-base-uncased version of BERT for binary classification [20]. Contextual embeddings generated by BERT are passed through a classification head to produce spam probability scores. The model is configured with a maximum sequence length of 128, batch size of 16, two training epochs, a learning rate of 2×10^{-5}, and the AdamW optimizer [23].

The model is trained in a GPU-enabled environment using the Hugging Face Transformers framework. Model performance is evaluated using accuracy, precision, recall, F1-score, and confusion matrix, with special attention to spam recall and false negatives because undetected spam creates higher security risk in practical use [25].

V. RESULTS AND DISCUSSION

This section compares the performance of traditional machine learning models and the transformer-based BERT model for SMS spam detection. The evaluation is based on accuracy, spam recall, spam F1-score, and confusion matrix results. Since undetected spam may lead to phishing and financial fraud, spam recall and false negatives are considered more important than accuracy alone.

A. Baseline Model Performance

Traditional classifiers were trained using TF-IDF features, and their performance is shown in Table I.

TABLE I- PERFORMANCE OF TRADITIONAL MACHINE LEARNING MODELS

MODEL	ACCURACY	SPAM RECALL	SPAM F-1 SCORE
Naïve Bayes	97.13%	0.79	0.88
Logistic regression	97.21%	0.79	0.88
Random Forest	97.76%	0.83	0.91
SVM	98.33%	0.90	0.94

As presented in Table I, all traditional models achieved high accuracy between 97.13% and 98.33%. SVM performed best among them, with 98.33% accuracy, 0.90 spam recall, and 0.94 spam F1-score. However, Naïve Bayes and Logistic Regression showed lower spam recall, indicating that some spam messages were still classified as legitimate. This shows that TF-IDF-based models are useful but limited in detecting context-dependent spam.

B. BERT Model Performance and Confusion Matrix

The fine-tuned BERT model achieved better performance than the traditional classifiers, with 99% accuracy, 0.97 spam recall, and 0.98 spam F1-score. It produced only 2 false positives and 5 false negatives.

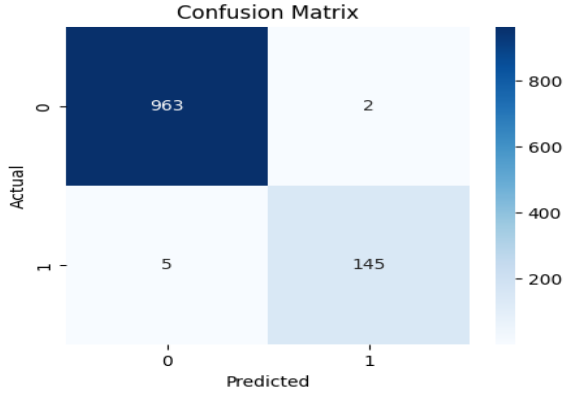

Fig. 1. Confusion matrix of the BERT model for SMS spam detection.

Fig. 1 shows that the BERT model correctly classified 963 legitimate messages and 145 spam messages. Only 2 legitimate messages were wrongly marked as spam, while 5 spam messages were missed. This low error rate shows that BERT can identify spam more accurately by understanding message context and semantic meaning.

Overall, the results show that traditional machine learning models provide a strong baseline, but BERT gives better spam recall and fewer false negatives. Therefore, transformer-based models are more suitable for real-time spam detection, especially where phishing and fraud messages are written in a short, indirect, or context-sensitive manner.

VI. CONCLUSION

This study proposed and evaluated a real-time AI-based spam detection system using transformer architecture for contextual message classification. The main objective was to compare traditional machine learning classifiers with a fine-tuned BERT model for SMS spam detection. Classical models such as Naïve Bayes, Logistic Regression, Random Forest, and Support Vector Machine achieved competitive accuracy values between 97% and 98%, but their spam recall was lower than that of the transformer-based model. The fine-tuned BERT model achieved 99% accuracy, 0.97 spam recall, and 0.98 spam F1-score, with very few false positives and false negatives. These findings show that transformer-based models provide stronger contextual understanding and semantic discrimination than traditional feature-based classifiers. Overall, the study supports the use of transformer-based spam detection systems for real-time communication environments where phishing, fraud, and large-scale spam campaigns are increasing.

VII. FUTURE WORK

Although the proposed transformer-based framework produced strong results on the SMS dataset, further work is required to make the system more suitable for real-world multi-channel deployment. Future research may extend the framework to a fully integrated system that processes SMS, email, social media comments, and instant messaging content in real time. Multilingual spam detection is another important direction because regional language use is increasing across Indian digital platforms. Future models may use multilingual transformer architectures to improve cross-language generalization. Adversarial robustness also needs further investigation, especially because AI-generated phishing messages are becoming more fluent and adaptive. Techniques such as adversarial training and continuous model updating may help the system remain effective against new spam patterns. Finally, low-latency deployment can be improved

using model quantization, knowledge distillation, and lightweight transformer architectures so that a better balance can be achieved between accuracy and computational efficiency.

REFERENCES

[1] Indian Computer Emergency Response Team (CERT-In), *Annual Report 2021*. New Delhi, India: Ministry of Electronics and Information Technology, Government of India, 2021.

[2] National Crime Records Bureau (NCRB), Crime in India Report 2021. New Delhi, India: Ministry of Home Affairs, Government of India, 2022.

[3] National Crime Records Bureau (NCRB), Crime in India Report 2023. New Delhi, India: Ministry of Home Affairs, Government of India, 2024.

[4] Press Information Bureau (PIB), "Cybercrime statistics report 2023," Government of India, New Delhi, India, 2024.

[5] Indian Computer Emergency Response Team (CERT-In), Cyber Security Incidents Report 2025. New Delhi, India: Ministry of Electronics and Information Technology, Government of India, 2025.

[6] Reserve Bank of India (RBI), Annual Report 2024–25. Mumbai, India: Reserve Bank of India, 2025.

[7] S. K. Sahu and B. K. Mishra, "A comprehensive survey on machine learning and deep learning techniques for spam detection," IEEE Access, vol. 9, 2021.

[8] A. Goswami, R. K. Patel, C. Mavani, and H. K. K. B. Mistry, "Identifying online spam using artificial intelligence," Int. J. Recent Innov. Trends Comput. Commun., vol. 12, no. 8, Aug. 2024.

[9] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, and C. D. Spyropoulos, "An evaluation of Naïve Bayesian anti-spam filtering," in Proc. Workshop on Machine Learning in Information Filtering, 2000.

[10] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian approach to filtering junk e-mail," in Proc. AAAI Workshop on Learning for Text Categorization, 1998.

[11] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.

[12] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[13] B. Klimt and Y. Yang, "Introducing the Enron corpus," in Proc. Conf. Email and Anti-Spam (CEAS), 2004.

[14] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, "Contributions to the study of SMS spam filtering," in Proc. ACM Symp. Document Engineering, 2011.

[15] A. Gupta, H. Lamba, and P. Kumaraguru, "Faking Sandy: Characterizing and identifying fake images on Twitter during Hurricane Sandy," in Proc. Int. World Wide Web Conf. (WWW), 2013.

[16] S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python. Sebastopol, CA, USA: O'Reilly Media, 2009.

[17] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge University Press, 2008.

[18] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed. Hoboken, NJ, USA: Pearson, 2023.

[19] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Inf. Process. Manage., vol. 24, no. 5, pp. 513–523, 1988.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2019.

[21] A. Vaswani et al., "Attention is all you need," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2017.

[22] L. Breiman, "Bagging predictors," Mach. Learn., vol. 24, no. 2, pp. 123–140, 1996.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learn. Representations (ICLR), 2015.

[24] F. Chollet, Deep Learning with Python. Shelter Island, NY, USA: Manning Publications, 2018.

[25] T. Fawcett, "An introduction to ROC analysis," Pattern Recognit. Lett., vol. 27, no. 8, pp. 861–874, 2006.

[26] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.